Indian Journal of Agricultural Sciences **87** (6): 754–8, June 2017/Article
https://doi.org/10.56093/ijas.v87i6.70960

Forecasting prices of coffee seeds using Vector Autoregressive Time Series Model

B S YASHAVANTH[1], K N SINGH[2], AMRIT KUMAR PAUL[3] and RANJIT KUMAR PAUL[4]

ICAR- Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 21 November 2014; Accepted: 18 January 2017

ABSTRACT

Forecasts of agricultural prices are useful to the farmers, policymakers and agribusiness industries. In this globalized world, management of food security in the developing countries like India where agriculture is dominated needs efficient and reliable price forecasting models. In the present study, Vector Autoregression (VAR) has been applied for modeling and forecasting of monthly wholesale price of clean coffee seeds in different coffee consuming centers, viz. Bengaluru, Chennai and Hyderabad. Augmented Dickey-Fuller (ADF) test has been used for testing the stationarity of the time series. The appropriate VAR model is selected based on minimum Akaike Information Criterion (AIC). The VAR model obtained is compared with the Auto Regressive Integrated Moving Average (ARIMA) models with respect to forecast accuracy measures. The residuals of the fitted models were diagnosed for possible presence of autocorrelation and Autoregressive Conditional Heteroscedasticity (ARCH) effects.

Kew words: AIC, ARIMA, Forecasting, Stationarity, VAR

Price forecasting is a vital part of commodity trading and price analysis. Prices of agricultural and horticultural commodities are highly varying as they are largely influenced by several eventualities. Natural calamities like droughts, floods and attacks by pests and diseases make these unpredictable leading to a considerable risk and uncertainty in the process of price modeling and forecasting. Forecasts of prices are intended to be useful to the farmers, governments, agribusiness industries and other stakeholders. They need internal forecasts to execute policies that provide technical and market support for the agricultural sector. Before liberalization and globalization, prices were controlled by the government, rendering price forecasting a low value-added activity. Presently, the domestic and international market forces determine the prices. This leads to increased price variability making it imperative to study the trends in prices of different commodities by employing sound statistical modeling techniques which in turn, will help the planners in formulating suitable policies to face the challenges ahead. The agricultural price forecasts are also important to farmers as it helps to strategize their production and marketing on the expected prices that may have financial repercussions many months later (Jha and Sinha 2013).

In time series modeling, the past observations of the same variable are collected and analyzed to develop a model describing the underlying relationship. During the past few decades, a lot of effort has been directed towards developing and improving time series forecasting models. One of the most important and widely used time series models is the Auto Regressive Integrated Moving Average (ARIMA) model. The popularity of ARIMA model is due to its statistical properties as well as use of well-known Box-Jenkins methodology in the model building process (Box *et al.* 2007). The ARIMA methodology has been used by several authors for agriculture related forecasting such as cultivated areas (Prabakaran *et al.* 2013), price (Assis *et al.* 2010), productions (Paul *et al.* 2014) and productivity (Padhan 2012) of different crops. Demand for tractors, transplanters and combine harvesters were also modeled using ARIMA methodology (Kim *et al.* 2013). If the seasonality is observed in the data, Seasonal ARIMA (Saz, 2011) can be made use of. However, when many series are to be modeled, ARIMA requires each series to be modeled separately consuming bundle of resources. One of the solutions is to go for multivariate time series analysis, like Vector autoregressive (VAR) models, where all the series are modeled at a go. Moreover it captures the relations between different series which helps in arriving at better models than those given by univariate time series models. Recent literature shows that VAR methodology has been used extensively for modeling economic variables. Sathianandan (2007) used VAR type of models to model and discover the relationships between landings of eight commercially important marine fish species/groups

[1]Ph D Scholar (e mail: yashavanthbs@gmail.com), [2]Principal Scientist and Head (e mail: knsingh@iasri.res.in), [3]Principal Scientist (e mail: pal@iasri.res.in), [4]Scientist (e mail: ranjitstat@iasri.res.in).

using quarterwise landings in Kerala during 1960-2005. Kilian (2011) forecasted the price of oil using Vector Autoregression. Trujiello-Barrera *et al.* (2013) forecasted hog prices in United States using VAR models. Gutierrez (2014) employed VAR methodology to analyze the world wheat market. The present study is directed at application of VAR time series model to forecast the monthly wholesale prices of clean coffee seeds in different coffee consuming centers. An attempt is also made to compare the results obtained with the ARIMA models.

MATERIALS AND METHODS

In an ARIMA model, time series variable is assumed to be a linear function of the past values and random shocks. In general, an ARIMA model is characterized by the notation ARIMA (p, d, q), where p, d and q denote orders of Auto-Regression (AR), Integration (differencing) and Moving Average (MA), respectively. ARIMA is a parsimonious approach which can represent both stationary and non-stationary processes.

An ARMA (p, q) process is defined by equation

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \quad (1)$$

where, y_t and ε_t are the actual value and random error at time period t, respectively, ϕ_i (i=1, 2,……,p) and ϕ_i (j=1, 2,……,q) are the model parameters. The random errors, ε_t are assumed to be independently and identically distributed with mean zero and variance σ^2 (Box *et al*. 2007).

The process of ARIMA modeling begins with checking the time series for stationarity as the estimation procedure is available only for a stationary series. A series is regarded stationary if its statistical characteristics such as the mean and the autocorrelation structures are constant over time. This can be achieved by differencing the series or going for transformations. After appropriate transformation and differencing, different ARMA models are chosen on the basis of Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) that closely fit the data. Then, the parameters of the tentative models are estimated through any of the non-linear optimization procedures such that the overall measure of errors is minimized or the likelihood function is maximized. Finally, diagnostic checking for model adequacy is performed for all the estimated models through the plot of residual ACF and using Portmanteau test. The most suitable ARIMA model is selected using the smallest Akaike Information Criterion (AIC) or Schwarz-Bayesian Criterion (SBC) value (Makridakis *et al*. 2003).

A univariate autoregression involves one variable. In a univariate autoregression of order p, we regress a variable on p lags of itself. In contrast, a multivariate autoregression, i.e., a vector autoregression, or VAR, involves k variables. In n k-variable VAR of order p, we estimate k different equations. In each equation, we regress the relevant left hand-side variable on p lags of it, and p lags of every other variable. Thus the right hand side variables are the same in every equation – p lags of every variable. The key point is that, unlike the univariate case, vector autoregressions allow for cross-variable dynamics. Each variable is related not only to its own past, but also to the past of all the other variables in the system.

Suppose there are k time series components $\{Y_{1t}\},\{Y_{2t}\},...,\{Y_{kt}\}$ for t=0,1,2,3,…,n at equally spaced time intervals. We can represent these components by a vector $Y_t = (Y_{1t}, Y_{2t}, ..., Y_{kt})^T$ which we call as a vector of time series. A vector time series with k components can be modeled by a vector autoregressive model of order p denoted by VAR(p), and its expression is

$$Y_t = \mu + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + ... + \beta_p Y_{t-p} + \varepsilon_t \quad (2)$$

μ is the mean vector of the series, β_i $(i=1,2,...,p)$ are $k \times k$ parameter matrices,

$$\varepsilon_t = (\varepsilon_{1t}, ..., \varepsilon_{kt})^T$$

are independently and identically distributed random innovation vectors having zero mean and constant dispersion matrix Σ.

The VAR modeling method is simple since all variables considered are endogenous and the Ordinary Least Square technique (OLS) can be applied for estimation of parameters making it advantageous over other multivariate modeling techniques like simultaneous equation models. (Gujarati *et al*. 2009).

The forecasting ability of different models is assessed with respect to two common performance measures, viz. the root mean squared error (RMSE) and the mean absolute percentage error (MAPE). The RMSE measures the overall performance of a model and is given by equation (3)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2} \quad (3)$$

where, y_t is the actual value for time t, $\hat{y}_t$ is the predicted value for time t, and n is the number of predictions. The second criterion, the mean absolute percentage error is a measure of average error for each point forecast and is given by equation (4)

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \times 100 \quad (4)$$

where the symbols have the same meaning as above. The model with least RMSE and MAPE values is considered as the best model for the data.

The monthly wholesale prices (per kilogram of clean coffee seeds) of Arabica coffee seeds in important coffee consuming centers, viz. Bengaluru, Chennai and Hyderabad were used for the study. The data covered a period of 159 months (January, 2001 to March, 2014). The first 149 data points were used for model fitting and the last 10 data points were used for model validation. The data were obtained from various issues of Coffee Data, published by Coffee Board, Government of India available at the website *www.indiacoffee.org*. The ADF test for stationarity and plotting were done using R programming language. The ARIMA and VAR models were fit using SAS 9.4 statistical package

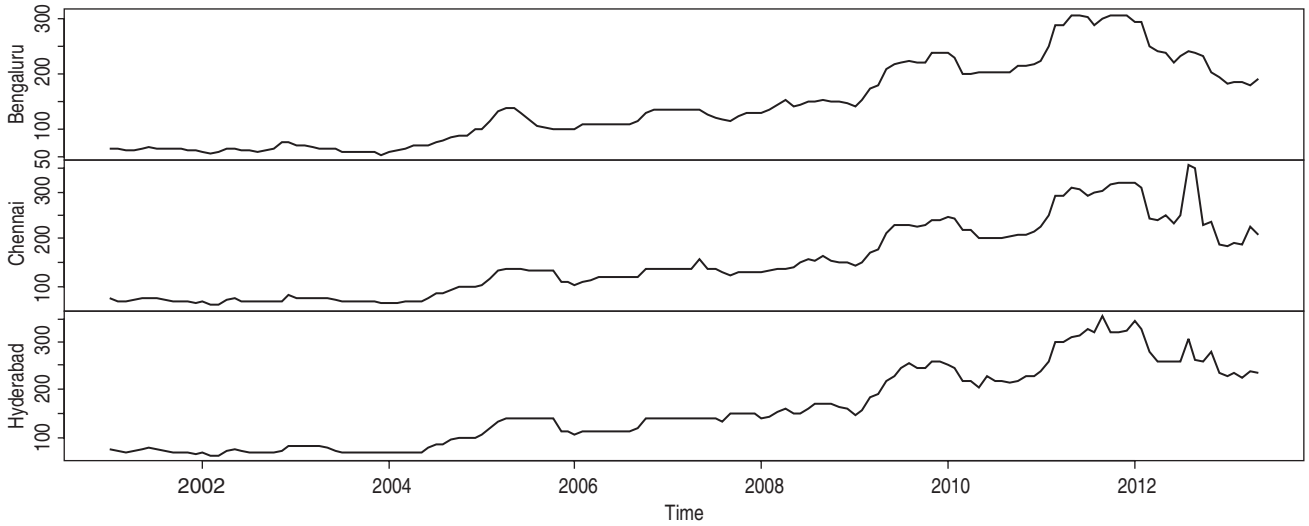

Fig 1 Monthly wholesale price of clean coffee seeds in different centers.

available at IASRI, New Delhi.

RESULTS AND DISCUSSION

The first step in time series analysis is to plot the data. Fig 1 shows the time series plot of monthly wholesale prices of clean seeds of arabica coffee for the period January 2001 to May 2013.

A perusal of Figure 1 reveals a positive trend over time which indicates the nonstationary nature of all the three time series. It can also be seen that all three series have similar behavioral pattern suggesting possible market cointegration in which changes occurring in one market results in changes in other markets. To confirm the presence of nonstationarity in the original data, a unit root test called Augmented Dickey-Fuller (ADF) test was applied, the results of which are given the Table 1. The Table 1 also includes the results of ADF test performed for all three series after first differencing. The values in Table 1 show the nonstationarity of all the three time series. Therefore, we have used first differencing for all three price series. The first differenced series were found to be stationary and hence further differencing was not required.

The ARIMA structure of differenced series is found out based on the autocorrelation function (ACF), partial autocorrelation function (PACF) and information criteria. We obtained the best ARIMA model for each series based on the lowest AIC and BIC information criteria. Among different candidate ARIMA models, we selected ARIMA (1,1,0) for Bengaluru series, ARIMA (7,1,0) for Chennai series. No candidate ARIMA models could be selected for the Hyderabad series based on the ACF and PACF plots, since none of the ACF and PACF values found significant. The models obtained are given in equations 5 and 6. The values in the parenthesis are the standard errors of the parameters.

Table 1 Results of Augmented Dickey-Fuller test for stationarity

Center	Original series		1st differenced series	
	ADF value	p - value	ADF value	p - value
Bengaluru	-2.576	0.337	-4.500	<0.01
Chennai	-2.226	0.482	-6.923	<0.01
Hyderabad	-2.835	0.228	-4.121	<0.01

For Bengaluru center,

$$\Delta y_{bt} = 0.351\ \Delta y_{bt-1} \quad (5)$$
$$(0.07787)$$

For Chennai center,

$$\Delta y_{ct} = -0.208\ \Delta y_{ct-1} - 0.298\ \Delta y_{ct-4} + 0.229\ \Delta y_{ct-7} \quad (6)$$
$$(0.084) \quad (0.083) \quad (0.086)$$

where Δy_{bt} and Δy_{ct} are the first differenced price values for Bengaluru and Chennai centers respectively, at time t.

VAR (p), were considered up to order p=5 and VAR (2) was selected as the best model based on the minimum AIC. The parameters of the model were obtained by least square estimation. For the VAR (2) model with 3 variables, initially 21 parameters were estimated and found that only 12 parameters were statistically significant at 90% confidence limits. Hence, the model parameters were estimated again by constraining the non significant elements to zero. The most striking point was that we could come up with an equation to forecast prices of the Hyderabad center for which the ARIMA model was not available.

The final models obtained for each of the series at 95% Confidence limits are given below.

For Bengaluru center,

$$\Delta y_{bt} = 1.224\ \Delta y_{bt-1} - 0.245\Delta\ y_{ct-2} \quad (7)$$
$$(0.034) \quad (0.033)$$

For Chennai center,

$$\Delta y_{ct} = 0.635\Delta\ y_{bt-1} - 0.496\Delta\ y_{ct-1} + 0.229\ \Delta y_{ht-1} - 0.363\ \Delta y_{ct-2} \quad (8)$$
$$(0.105) \quad (0.085) \quad (0.100) \quad (0.062)$$

For Hyderabad center,

$$\Delta y_{ht} = 0.739\ \Delta y_{bt-1} - 0.150\ \Delta y_{ct-1} + 0.610\ \Delta y_{ht-1} \quad (9)$$
$$(0.094) \quad (0.055) \quad (0.071)$$

where Δy_{ht} is the first differenced price value for Hyderabad center at time t.

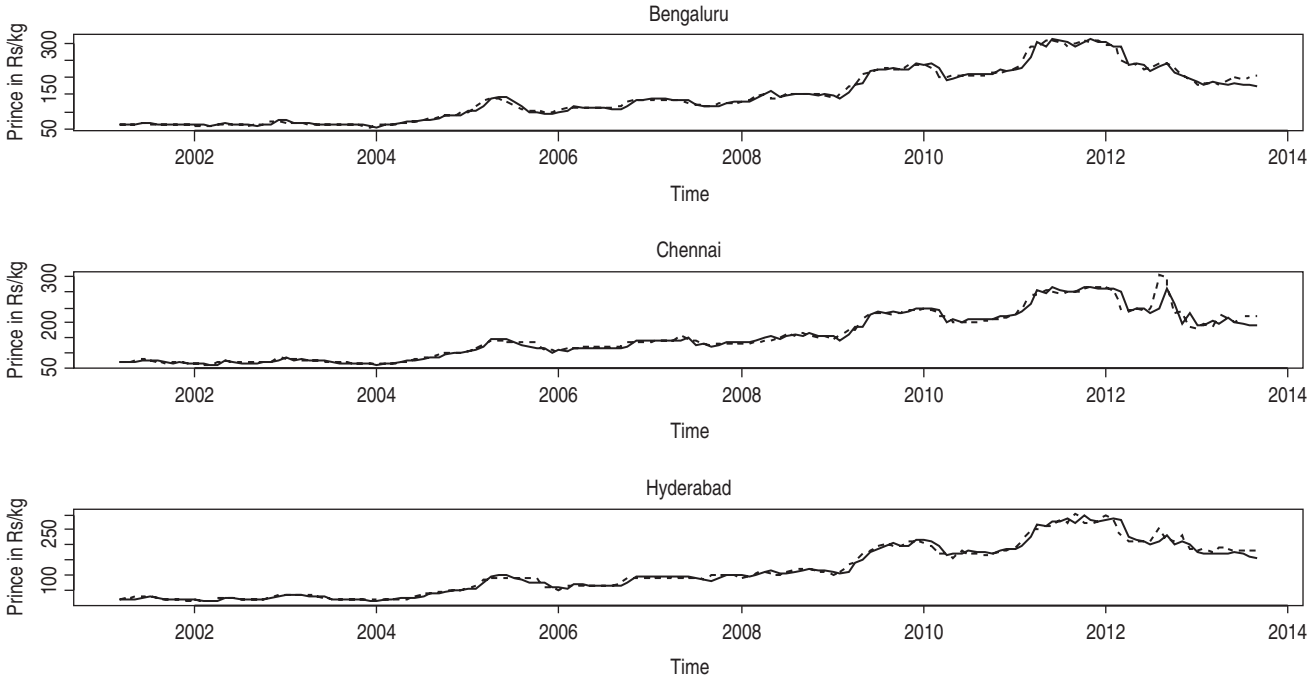

Fig 2 Actual and forecasted monthly wholesale price of clean coffee seeds in different centers.

Table 2 Individual model white noise diagnostics

Variable	Autocorrelation (Durbin-Watson statistics)	ARCH effect	
		F value	p > F
Bengaluru, yb	1.924	0.92	0.338
Chennai, yc	1.953	2.62	0.108
Hyderabad, yh	1.950	2.25	0.136

Table 3 Forecasting performance of different models for different centers

Center	RMSE		MAPE (%)	
	ARIMA	VAR	ARIMA	VAR
Bengaluru	8.498	8.081	3.754	3.723
Chennai	15.756	14.189	5.111	4.671
Hyderabad		10.592		4.114

These individual models fitted through VAR (2) were found highly significant (with F-values 1875.86, 640.81 and 1285.14 for Bengaluru, Chennai and Hyderabad centers, respectively). From the models obtained, we can also say that the present and future values of a series are not only dependent on the previous values of that particular series, but also on the previous values of other series in the system.

The model verification is concerned with checking residuals of the model to see if there are autocorrelation and ARCH effects. Table 2 describes how well each univariate equation fits the data. The Durbin-Watson statistics are close to 2.0 indicating the absence of autocorrelation between the residuals. Similarly, the nonsignificant F-values indicate the absence of ARCH effects.

The comparative results for the ARIMA and VAR models with respect to RMSE and MAPE for different series are given in Table 3. From the table, it can be seen

Table 4 The forecasted monthly prices (in Rs/kg) and their standard errors for different centers

Obs. No.	Month & Year	Bengaluru		Chennai		Hyderabad	
		Price	SE	Price	SE	Price	SE
160	Apr,2014	202.01	44.96	223.00	59.33	256.55	46.20
161	May,2014	202.91	47.23	223.99	62.06	257.73	48.46
162	Jun,2014	203.80	49.39	224.95	64.67	258.93	50.63
163	Jul,2014	204.69	51.47	225.93	67.18	260.13	52.71
164	Aug,2014	205.59	53.46	226.91	69.60	261.33	54.71
165	Sep,2014	206.48	55.38	227.89	71.94	262.53	56.64
166	Oct,2014	207.38	57.24	228.87	74.20	263.72	58.50
167	Nov,2014	208.27	59.04	229.85	76.40	264.92	60.31
168	Dec,2014	209.17	60.78	230.82	78.54	266.12	62.06
169	Jan,2015	210.06	62.48	231.80	80.62	267.32	63.77

that the both MAPE and RMSE values of VAR models are less than those of ARIMA models for both Bengaluru and Chennai center price series.

The one-step ahead forecasts for the monthly wholesale prices of clean Arabica coffee seeds, along with their standard errors, in three different centers using VAR (2) model are reported in Table 4. The result indicates that the prices are going to increase steadily with Hyderabad center reporting the highest price and Bengaluru center the least. The high values for standard errors signify that the prices are going to fluctuate drastically.

Figure 2 shows the plot of actual (dotted lines) and VAR forecasted (solid lines) monthly wholesale prices of clean seeds of arabica coffee in different markets. From the plots, it is conspicuous that the fitted values for price are close to the actual price values.

In this paper a multivariate time series modeling technique called Vector autoregression (VAR) has been used to model and forecast the monthly wholesale prices of arabica coffee in three different coffee consuming centers, viz. Bengaluru, Chennai and Hyderabad. The VAR model arrived is also compared with the univariate ARIMA models and it was found that VAR models fitted better than ARIMA models with respect to forecast accuracy measures and standard errors of the forecasted values. It was also evident that when an ARIMA model is not available for modeling a series, one can go for VAR model which makes use of the information available in other series when the series are cointegrated.

REFERENCES

Assis K, Amran A and Remali Y. 2010. Forecasting cocoa bean prices using univariate time series models. *Journal of Arts Science and Commerce* **1**(1): 71–80.

Box G E P, Jenkins G M and Reinsel G C. 2007. *Time-Series Analysis: Forecasting and Control*, 3rd edition. Pearson education, India.

Diebold, F.X. 2004. *Elements of Forecasting*. Thomson South – Western. India.

Gujarati D N, Porter D C and Gunasekar S. 2009. *Basic Econometrics*. Tata McGraw-Hill, New Delhi.

Jha G K and Sinha K. 2013. Agricultural price forecasting using neural network model: An innovative information delivery system. *Agricultural Economics Research Review* **26**(2): 229–39.

Kilian L. 2011. Real-time forecasts of the real price of oil. *Journal of Business and Economic Statistics* **30**(2): 326–36.

Makridakis S, Wheelright S C and Hyndman R J. 2003. *Forecasting: Methods and Applications*. Wiley-India. New Delhi.

Padhan PC. 2012. Application of ARIMA model for forecasting agricultural productivity in India, *Journal of Agriculture and Social Sciences* **8**: 50–6.

Paul R K, Alam W and Paul A K. 2014. Prospects of livestock and dairy production in India under time series framework. *Indian Journal of Animal Sciences* **84**(4): 130–4.

Prabakaran K, Sivapragasam C, Jeevapriya C and Narmatha A. 2013. Forecasting cultivated areas and production of wheat in India using ARIMA model, *Golden Research Thoughts* **3**(3).

Sathianandan T V. 2007. Vector time series modeling of marine fish landings in Kerala. *Journal of the Marine Biological Association of India* **49**(2): 197–205.

Saz G. 2011. The efficacy of SARIMA models for forecasting inflation rates in developing countries: The case for Turkey. *International Research Journal of Finance and Economics* **62**: 111–42.

Trujillo-Barrera A, Garcia P and Mallory M. 2013. Price density forecasts in the US hog market: Composite Procedures. *Proceedings of the NCCC-134 conference on Applied Commodity Price Analysis, Forecasting and Market Risk Management*, St. Louis, MO.*www.indiacoffee.org/database-coffee.html.*